TEST Engineering & Management

May-June 2020
ISSN: 0193-4120 Page No. 3529-3534

Machine Learning Algorithms for Big Scholarly Data: A Novel Approach

Raghavendra Nayaka P[1#], Rajeev Ranjan[2*]
[#1]School of C& IT, REVA University, Bangalore, India, raghavendranayak@reva.edu.in
[*2]School of CSA, REVA University, Bangalore, India, rajeevranjan@reva.edu.in

Article Info
Volume 83
Page Number: 3529-3534
Publication Issue:
May-June 2020

Article History
Article Received: *19 August 2019*
Revised: *27 November 2019*
Accepted: *29 January 2019*
Publication: *12 May2020*

Abstract

With the fast growth of publishing the research data, managing and analyzing the scholarly data has been challenging for the researchers. The term big scholarly data contains all the information including several research papers published by various authors over the globe, these research papers includes citations ,figures, tables etc., as well as scholarly networks and digital libraries. This paper presents a on how the techniques of the machine learning can be applied over Big Scholarly data which is much needed work for maintain quality and analysis over scholarly data. This review paper also provides a critical analysis of the work on what exactly this Big Scholarly Data is, by comparing with various methods along with existing research problems. Then the review focuses on Machine learning techniques, framework comparison and various algorithm used in the existing research work. Finally the works concentrates on applications of Machine Learning over Big Scholarly Data, along with the challenges faced in the existing work with how Machine Learning techniques can be applied over Big Scholarly Data and can be processed further is discussed.

Keywords: Big Scholarly Data; Machine Learning Challenges; Machine Learning Methods; Big Scholarly Data

1. Introduction

As much application run on web based the importance Big Scholarly Data has grown rapidly and also cannot keep track of all the this data so needs to be trained using machine learning algorithms to train the data and later can be used for analyzing the big scholarly data.

Both the Big Data and Machine learning techniques can be applied for analyzing the scholarly information. For every application can produce information. This circumstance will turn out to be more terrible if every gadget can be associated with different gadgets to utilize their data. At the end of the day, with the development of Internet of Things are looking with gigantic measure of information that should have been put away and overseen one of the example for big scholarly data is nutshell, where the advances in computerized gadgets are presented. For example, advanced sensors, a lot of information have been produced at a quick speed that brought about a region named Big Scholarly Data. Enormous Scholarly Data isn't just about delivering information from sensors; It can be given by people, writings, pictures et cetera.

Huge Scholarly Data greatly affects innovations and figuring. As such, this research has more information nowadays that present strategies can't manage this information. In straightforward approach term of Scholarly data implies gathering, preparing and exhibiting the aftereffects of tremendous measures of the data that comes from various sources.

In any case, ideally it needs high volumes of information. On the off chance that needs to be ended up more fruitful in this focused zone, we have to discover unique examples. The more achievement can be done by giving more number of examples In such cases the Machine Learning for Big Scholarly Data can be applied, based on the utilization user's commented as:

- A convenient prologue to Big Scholarly Data and looking at its strategies.
- A convenient prologue to Machine getting the hang of, looking at calculations and structures.

This paper describes review on the big scholarly data and its challenges, The section 3 describes Machine Learning techniques over scholarly data, the section 4 describes how Machine Learning for future patterns, and

May-June 2020
ISSN: 0193-4120 Page No. 3529-3534

open research issues. Finally paper concluded with proper conclusion and the future enhancement

2. Review on Big Scholarly Data

The growth of the scholarly data has been increasing because of various publications and the research done in the different fields which causes the capability of the data storage, increase in the computations power utilities and the more accessibility of data volume. The most important part of existing works is to deal with analyses of scholarly data is a challenging task which is mainly concentrating on various issues like the V-factor which includes volume, velocity, validity, variety and the volatility.

The first factor is the Volume that defines how the colossal measures of the information that a large portion of conventional calculations are not ready to manage this test.

The next is validity defines how spotless, reliable, handiness; result information ought to be substantial, as feasible for later handling stages. Next V's describes the variety and the volatility of the scholarly data defines the size and the sources taken for analyzing the scholarly data[6].

Table 1: Comparison between machine learning techniques

Algorithm	Type of Learning	Class	Restriction Bias	Preference Bias
K- Nearest Neighbor [2]	Supervised Learning	Instance Based	Works well for measuring the distance between approximations	Prefers problems which are distance based
Naïve Byes [4]	Supervised Learning	Probabilistic	Works on the problems where the input dependent on other variables	Prefers problems where probability is always zero class
Hidden Markov Chain [5]	Supervised Learning	Markovian	Works well on Markova assumption models	Prefers memory classification problems
Support Vector machine [7]	Supervised Learning	Decision Boundary	Works for distinction based classes	Prefers binary classification problems
Neural Network [8]	Supervised Learning	Non-Linear functional Approximation	Has little restriction bias	Prefers only binary inputs
Clustering [9]	Un-Supervised Learning	Clustering	No restriction on the bias	Prefers data only grouping based
Regression [11]	Supervised Learning	Linear Regression	Low restriction	Prefers only on continuous variables
Filtering [12]	Un-Supervised Learning	Transformation of the features	No restriction	Prefers data lot of variables in the filter

Labrinidis et.al [7] describes a few difficulties in the existing research issues as for Scalability, Heterogeneity parts of Big Scholarly Data administration. Different parameters, for example, accessibility and honesty are shrouded in [8]. The parameters considered for this analysis are the availability and the reliability Means the information taken to be open and accessible at whatever point and wherever client demands information even on account of disappointment event. Information investigation strategies Ought to give accessibility to help a lot of information alongside a rapid stream of 2011 for mechanical applications to scale well in constrained memory.

❖ **Data Integrity:** focuses to information precision. The circumstance be-comes more terrible when diverse clients with various benefits change information in the cloud. Cloud is responsible for overseeing databases. Along these lines, clients need to obey cloud strategy for information uprightness [10].

❖ **Resource Optimization:** implies utilizing existing assets effectively. An exact strategy for asset streamlining is required for ensuring dispersed access to Big Scholarly Data.

A. Preprocessing the Data

For better basic leadership, the quality information has been given to information dissecting step. As it were, the nature of information is basic to quality choice. Likewise the information has to be confirmed before choice. Preprocessing information implies changing, in-consistency, fragmented information that has numerous mistakes into a suitable configuration for additionally investigations. As it were, information must be organized before examination arranges [13]. There are a few stages

May-June 2020
ISSN: 0193-4120 Page No. 3529-3534

for accomplishing preprocessing area objective as depicted as takes after:

- ❖ Purifying the data: Removing errors, inadequacy, and irregularities of information.
- ❖ Data change: changing the data means doing extra procedures like total, or change. This progression affects future advances.
- ❖ Data coordination: It gives a solitary view over appropriated information from various sources.
- ❖ Data transmission: Defines a strategy for exchanging crude information to capacity framework, for example, question stockpiling, server farm or circulated distributed storage.
- ❖ Data decrease: diminishing the span of substantial databases for ongoing applications [14]. The accompanying sub-segments display more insight about some preprocessing steps:
- ❖ Data Cleansing: In basic word implies distinguishing deficient and silly information. This can be altered or erased these sorts of information with a specific end goal to accomplish quality change for additionally preparing advances. Maletic and Marcus mulled over five phases with a specific end goal to accomplish clean information: 1) perceiving sorts of mistakes 2) discovering blunder occurrences 3) redress mistake occasions and blunder composes 4) refresh information input method keeping in mind the end goal to decrease advance blunders that may happen 5) checking information issues like confinements, arrangements, and rationalities. Information purging is an irreplaceable and chief piece of information examination step.
- ❖ Mobility of the data: It defines the quantity of steps that are required to get the last outcome.
- ❖ Partitioning of the data: The calculations utilized for parceling information. In concise of apportioning procedures used to be utilizing keeping in mind the end goal to accomplish better information parallelism.
- ❖ Data Availability: Data availability presents a method that ensures information accessibility if there should occur an occurrence of disappointments event.

B. Data Storage

Putting away information in pet byte scale is a test for scientists as well as for web associations. Nowadays we can barely adjust existing databases to Big Scholarly Data use. Despite the fact that Cloud Computing uncovers a move to another figuring worldview, it can't guarantee consistency effortlessly while putting away Big Scholarly Data in distributed storage. It's anything but a decent method to squander information since it might add to better basic leadership. So it is basic to have a capacity administration framework keeping in mind the end goal to give enough information stockpiling, and advanced data recovery [13].

Replication: Replication is a major action that makes information accessible and available at whatever point client inquires [16].

Ordering: For expansive databases, it isn't insightful to recover put away information and looking information in consecutive frame like an un-requested exhibit [17]. Ordering information enhances the execution of capacity administrator. So proposing a reasonable ordering instrument is testing.

C. Big Scholarly Data Processing and Management

There are mainly four different types of information models that has been concentrating in Big Scholarly Data where 1st zone describes the information that can store them in social 2-semi-organized information same as XML 3-chart information, for example, those who use for online networking and the last one is unstructured information, for example, content information, transcribed articles [21].

3. Machine Learning and Challenges

A. Machine Learning

By and large, there are kinds of machine learning where the first is Shallow learning, for example Support Vector Machines (SVMs) that it is probably going to miss the mark whenever there is a need to extricate valuable data from gigantic measures of information and regardless of whether they would not miss the mark, they won't have fulfilled precision.

A vital inquiry here is with all the distinctive calculations in the ML, how might the pick for the best one for our motivation? On the off chance that needs to anticipate or gauge an objective esteem, at that point should utilize administered learning procedures, for example, Neural Networks (NN) that can be known for the right answers beforehand. At the end of the day, managed learning issues are sorted into "relapse" and "arrangement" issues. In a relapse issue, there are attempting to anticipate yield of consistent qualities, implying that we are endeavoring to delineate factors to some ceaseless capacities [32]. At the end of the day, endeavoring to delineate factors into discrete classes [33] has to be done.

4. Use of Machine Learning over Big Scholarly Data

This section describes how the uses of machine learning techniques over big data can be applied along with the applications are discussed [36].

A. Machine Learning techniques for Larger data

The first step is to check how much scholarly data can be analyzed. When all is said in done, then apply DL calculations in a segment of accessible Big Scholarly Data for preparing objective and utilize whatever is left of information for separating conceptual portrayals and from another perspective, question is that how much volume of information is required for preparing information.

May-June 2020
ISSN: 0193-4120 Page No. 3529-3534

Another open issue is area adjustment, in applications which preparing information is not the same as the dispersion of test information.

Another issue is characterizing criteria for enabling information portrayals to give valuable future semantic implications. In basic word, each removed information portrayal ought not to be permitted to give valuable importance.

Another is that a large portion of the DL calculations require a predefined misfortune and that should be recognized by what is our mean to separate, in some cases it is exceptionally hard to comprehend them in the Big Scholarly Data condition.

The other issue is that the vast majority of them don't give logical outcomes that can be reasonable effortlessly. As such, in light of its unpredictability, you can't break down the system effectively. This circumstance turns out to be more terrible in a Big Scholarly Data condition.

The last however not the slightest real issue is that they require named information. On the off chance that cannot be given as marked information, they will have terrible execution. One conceivable answer for this is can be utilized for fortification taking in, the framework accumulates information without anyone else's input, and the main requirement for us is offering prizes to the framework.

B. Machine Learning for High Variety of Data

Nowadays the information comes from wide range of arrangements from various sources, presumably with various circulations. For instance, the quickly developing interactive media information originating from the web and cell phones incorporate a gigantic gathering of pictures, recordings and sound streams, designs and movements, and unstructured content, each with various qualities. There are open inquiries in such manner that should be tended to as some of them displayed as takes after:

- Given that diverse sources may offer clashing data, how might can be settled the contentions and circuit the information from various sources adequately and productively?
- if the framework execution profits by essentially developed modalities?
- in which level Machine Learning designs are proper for include combination of heterogeneous information?

C. Machine Learning for High Velocity of Data

Information is producing at to a great degree rapid and should be handled at quick speed. One answer for gaining from such high-speed information is web based learning approaches that should be possible by Machine learning. Just restricted advancement in online Machine Learning has been made as of late.

5. Results and discussion

The experimental results can be drawn based on various open source resources available in web from various journals, For the fetched Scholar APIs, the class name of attribute must be specified.

The details that are fetched from Scholarly based application are stored in CSV format that contains various attributes of the abstract, introduction, methodology, keywords, conclusion etc. That forms five (5) different types of clusters for the attributes taken. Then, the formed clusters will be stored in separate files, which can be used for classification purpose.

Table 2: Number of instances considered in existing works

Number of journals instances	**50**	**100**	**200**	**300**
Accuracy of existing works done	35	45	55	65

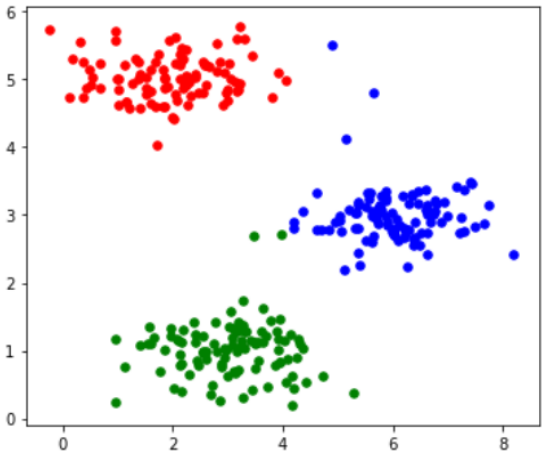

Figure 2: The Sample representation of Clusters

From the Scholarly information dataset taken with various instances have been taken and grouped into three clusters. Below is tabulation n showing the distribution of instances for each cluster**.**

From the above information we drawn various results and their accuracy by applying various improvised machine leaning techniques and big data techniques, we try to improve the algorithm and prediction accuracy in the proposed work.

6. Conclusion

Nowadays, it's necessary to collect all the necessary knowledge with the aim of extracting abstract information from each scholar paper. One technique that's applicable for this aim is Machine Learning that has higher-level knowledge abstraction. Machine Learning could be a helpful technique that may conjointly be used in the huge erudite knowledge atmosphere and has its own benefits and downsides. In general, the lot of knowledge, the upper level abstract knowledge, however we tend to face several challenges. This paper shows

May-June 2020
ISSN: 0193-4120 Page No. 3529-3534

initially huge erudite knowledge steps, then Machine learning and Machine Learning and finally application of Machine Learning in huge erudite knowledge, future trends, and open analysis issues. In the future, this work has got a thought to listen to higher than areas in additional detail and conjointly work huge erudite knowledge issues within the trade. This tends to area unit progressing to even have a survey on huge erudite knowledge security and privacy issue. Then would like to deal with different issues like linguistics compartmentalization, knowledge tagging and then on.

References

[1] Yoshua Bengio et al. Deep Learning deep architectures for ai. Foundations and trends R in Machine Learning, 2(1):1–127, 2018.

[2] Jinchuan Chen, Yueguo Chen, Xiaoyong Du, Cuiping Li, Jiaheng Lu, Suyun Zhao, and Xuan Zhou. Big Scholarly Data challenge: KNN a data management perspective. Frontiers of Computer Science, 7(2):157–164, 2018.

[3] Dylan Maltby. Big Scholarly Data analytics. In 74th Annual Meeting of the Association for Information Science and Technology (ASIST), pages 1–6, 2018.

[4] Han Hu, Yonggang Wen, Tat-Seng Chua, and Xuelong Li. Toward scalable systems for Big Scholarly Data analytics: A technology tutorial using Naïve Bayes technique. IEEE Access, 2:652–687, 2017.

[5] Aisha Siddiqa, Ibrahim Abaker Targio Hashem, Ibrar Yaqoob, Mohsen Marjani, Shahabuddin Shamshirband, Abdullah Gani, and Fariza Nasaruddin. A Hidden Markov Chain for Big Scholarly Data management: taxonomy and state-of- the-art. Journal of Network and Computer Applications, 71:151–166, 2016.

[6] Min Chen, Shiwen Mao, and Yunhao Liu. Big Scholarly Data: A survey. Mobile Networks and Applications, 19(2):171–209, 2014.

[7] Alexandros Labrinidis and Hosagrahar V Jagadish. Challenges and opportunities with Big Scholarly Data. Using Support Vector machine, Proceedings of the VLDB Endowment, 5(12):2032–2033, 2015.

[8] Chang Liu, Chi Yang, Xuyun Zhang, and Jinjun Chen. External integrity verification for outsourced Big Scholarly Data Neural Network in cloud and iot: A big picture. Future Generation Computer Systems, 49:58–67, 2015.

[9] Katina Michael and Keith W Miller. Big Scholarly Data: New opportunities Clustering techniques guest editors' introduction. Computer, 46(6):22–24, 2013.

[10] Ian H Witten, Eibe Frank, Mark A Hall, and Christopher J Pal. Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann, 2016.

[11] Xue-Wen Chen and Xiaotong Lin. Big Scholarly Data Machine learning using liner Regression challenges and perspectives. IEEE Access, 2:514–525, 2016.

[12] CL Philip Chen and Chun-Yang Zhang. Data-intensive using Filtering, techniques and technologies: Data. Infor- mation Sciences, 275:314–347, 2014.

[13] Andrew McAfee, Erik Brynjolfsson, Thomas H Davenport, DJ Patil, and Dominic Barton. Big Scholarly Data. The management revolution. Harvard Bus Rev, 90(10):61–67, 2012.

[14] Geoffrey E Hinton and Ruslan R Salakhutdinov. Reducing the dimen- sionality of data with neural networks. Science, 313(5786):504–507, 2016.

[15] Anton Riabov and Zhen Liu. Scalable planning for distributed stream processing systems. In ICAPS, pages 31–41, 2017.

[16] J u¨rgen Schmidhuber. Deep learning in neural networks: An overview. Neural networks, 61:85–117, 2015.

[17] Jens Dittrich, Lukas Blunschi, and Marcos Antonio Vaz Salles. Movies: indexing moving objects by shooting index images. Geoinformatica, 15(4):727–767, 2011.

[18] Guoliang Li, Beng Chin Ooi, Jianhua Feng, Jianyong Wang, and Lizhu Zhou. Ease: an effective 3-in-1 keyword search method for unstructured, semi-structured and structured data. In Proceedings of the 2008 ACM SIGMOD international conference on Management of data, pages 903– 914. ACM, 2008.

[19] Li Deng, Dong Yu, et al. Machine learning: methods and applications. Foundations and Trends§R in Signal Processing, 7(3–4):197–387, 2014.

[20] Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. Extracting and composing robust features with denoising autoencoders. In Proceedings of the 25th international conference on Machine learning, pages 1096–1103. ACM, 2008.

[21] Divyakant Agrawal, Amr El Abbadi, Shyam Antony, and Sudipto Das. Data management challenges in cloud computing infrastructures. In International Workshop on Databases in Networked Information Systems, pages 1–10. Springer, 2010.

[22] GNU Octave. Gnu octave. l´ınea]. Available: http://www. gnu. org/software/octave, 2012.

[23] Xiao Chen. Google big table. 2010.

[24] Da-Wei Sun, Gui-Ran Chang, Shang Gao, Li-Zhong Jin, and Xing-Wei Wang. Modeling a dynamic data replication strategy to increase system availability in cloud computing environments. Journal of computer science and technology, 27(2):256–272, 2012.

May-June 2020
ISSN: 0193-4120 Page No. 3529-3534

[25] Daniel E O'Leary. Artificial intelligence and Big Scholarly Data. IEEE Intelligent Systems, 28(2):96–99, 2013.

[26] Vivien Marx. Biology: The big challenges of big data. Nature, 498(7453):255–260, 2013.

[27] P. Porkar. Sensor networks challenges. In 11th international conference on data networks, DNCOCO '12,, 7-9 September 2012.

[28] Xavier Glorot, Antoine Bordes, and Yoshua Bengio. Domain adaptation for large-scale sentiment classification: A Machine Learningapproach. In Proceedings of the 28th international conference on machine learning (ICML-11), pages 513–520, 2011.

[29] Shahabi Amir. clustering algorithm in Wireless Sensor network, chapter Sustainable Interdependent Networks: From Theory to Application. Springer, accepted for publication (2018).

[30] Krisztian Buza, G a´bor I. Nagy, and Alexandros Nanopoulos. Storage- optimizing clustering algorithms for high-dimensional tick data. Expert Syst. Appl., 41:4148–4157, 2014.

[31] Mehdi Jafari, Jing Wang, Yongrui Qin, Mehdi Gheisari, Amir Shahab Shahabi, and Xiaohui Tao. Automatic text summarization using fuzzy inference. In Automation and Computing (ICAC), 2016 22nd Interna- tional Conference on, pages 256–260. IEEE, 2016.

[32] Dervis Karaboga and Celal Ozturk. A novel clustering approach: Artifi- cial bee colony (abc) algorithm. Applied soft computing, 11(1):652–657, 2011.

[33] Mehdi Gheisari, Ali Akbar Movassagh, Yongrui Qin, Jianming Yong, Xiaohui Tao, Ji Zhang, and Haifeng Shen. Nsssd: A new semantic hierarchical storage for sensor data. In Computer Supported Cooperative Work in Design (CSCWD), 2016 IEEE 20th International Conference on, pages 174–179. IEEE, 2016.

[34] Surajit Chaudhuri, Umeshwar Dayal, and Vivek Narasayya. An overview of business intelligence technology. Communications of the ACM, 54(8):88–98, 2011.

[35] T. Tran, M. Rahman, M. Z. A. Bhuiyan, A. Kubota, S. Kiyomoto, and K. Omote. Optimizing share size in efficient and robust secret sharing scheme for Big Scholarly Data. IEEE Transactions on Big Scholarly Data, PP(99):1–1, 2017.

[36] Richard S Sutton and Andrew G Barto. Introduction to reinforcement learning, volume 135. MIT Press Cambridge, 1998.

[37] Steve Lohr. The age of Big Scholarly Data. New York Times, 11(2012), 2012.

[38] M. Z. A. Bhuiyan and J. Wu. Event detection through differential pattern mining in cyber-physical systems. Jun 2017.

[39] W. Yu, J. Li, M. Z. A. Bhuiyan, R. Zhang, and J. Huai. Ring: Real- time emerging anomaly monitoring system over text streams. IEEE Transactions on Big Scholarly Data, PP(99):1–1, 2017.

[40] Katherine G Herbert and Jason TL Wang. Biological data cleaning: a case study. International Journal of Information Quality, 1(1):60–82, 2007.

[41] Maryam M Najafabadi, Flavio Villanustre, Taghi M Khoshgoftaar, Naeem Seliya, Randall Wald, and Edin Muharemagic. Machine Learningapplications and challenges in Big Scholarly Data analytics. Journal of Big Scholarly Data, 2(1):1, 2015.

[42] Todd A Letsche and Michael W Berry. Large-scale information retrieval with latent semantic indexing. Information sciences, 100(1-4):105–137, 1997.

[43] Vijay Khatri and Carol V Brown. Designing data governance. Communications of the ACM, 53(1):148–152, 2010.